# Bayesian Models for Pooling Microarray Studies with Multiple Sources of Replications

BMC Bioinformatics

Methodology article



# Bayesian models for pooling microarray studies with multiple sources of replications

Erin M Conlon*[1], Joon J Song[1] and Jun S Liu[2]

Address: [1]Department of Mathematics and Statistics, University of Massachusetts, Amherst, Massachusetts, USA and [2]Department of Statistics, Harvard University, Cambridge, Massachusetts, USA

Email: Erin M Conlon* - econlon@mathstat.umass.edu; Joon J Song - jjsong@uark.edu; Jun S Liu - jliu@stat.harvard.edu

* Corresponding author

Published: 05 May 2006

*BMC Bioinformatics* 2006, **7**:247 doi:10.1186/1471-2105-7-247

Received: 03 August 2005
Accepted: 05 May 2006

This article is available from: http://www.biomedcentral.com/1471-2105/7/247



## Abstract

**Background:** Biologists often conduct multiple but different cDNA microarray studies that all target the same biological system or pathway. Within each study, replicate slides within repeated identical experiments are often produced. Pooling information across studies can help more accurately identify true target genes. Here, we introduce a method to integrate multiple independent studies efficiently.

**Results:** We introduce a Bayesian hierarchical model to pool cDNA microarray data across multiple independent studies to identify highly expressed genes. Each study has multiple sources of variation, i.e. replicate slides within repeated identical experiments. Our model produces the gene-specific posterior probability of differential expression, which provides a direct method for ranking genes, and provides Bayesian estimates of false discovery rates (FDR). In simulations combining two and five independent studies, with fixed FDR levels, we observed large increases in the number of discovered genes in pooled versus individual analyses. When the number of output genes is fixed (e.g., top 100), the pooled model found appreciably more truly differentially expressed genes than the individual studies. We were also able to identify more differentially expressed genes from pooling two independent studies in *Bacillus subtilis* than from each individual data set. Finally, we observed that in our simulation studies our Bayesian FDR estimates tracked the true FDRs very well.

**Conclusion:** Our method provides a cohesive framework for combining multiple but not identical microarray studies with several sources of replication, with data produced from the same platform. We assume that each study contains only two conditions: an experimental and a control sample. We demonstrated our model's suitability for a small number of studies that have been either pre-scaled or have no outliers.

## Background

cDNA microarrays monitor gene expression for thousands of genes simultaneously. Two experimental conditions are compared by examining the ratio of expression between two samples, e.g. treatment versus control, wildtype versus mutant, or disease versus healthy. The primary goal of these experiments is to identify genes that are differentially expressed between the two conditions. The up-regulated and down-regulated genes shed light on biological

mechanisms of the cell, such as functional pathways, response to treatments, and gene regulation.

Bayesian models have been used extensively in single microarray studies to identify differentially expressed genes (e.g. [1-3]); the discrete mixture model approach in particular has had considerable use [4-13]. Additional Bayesian methods for identifying differentially expressed genes include Bayesian ANOVA for microarrays (BAM) of Ishwaran and Rao [14,15]. This approach redefines the search for differentially expressed genes as a Bayesian variable selection process and uses a hierarchical model that is tailored to adaptive shrinkage. Through the use of model averaging, BAM essentially shrinks only the effects of the non-differentially expressed genes relative to the least squares estimates. For a review of Bayesian approaches to microarray data analysis and their advantages over frequentist methods, see Yang *et al.* [16].

In addition to performing single microarray studies, biologists often conduct multiple but not identical studies to understand the same biological system. Pooling results of these studies can help identify truly differentially expressed genes. Meta-analyses for microarray studies have been used recently by many researchers in a non-Bayesian context [17-24]. Rhodes *et al.* [17] combined the results of four prostate cancer studies using a *p*-value approach. Genes were assigned a *p*-value in each study separately, and the results were combined to estimate a gene-specific *p*-value across all studies. This method avoids the necessity of integrating gene expression measures and thus can be used for data across multiple platforms. Choi *et al.* [18] presented a meta-analysis that integrated gene effect sizes, rather than *p*-values, into one mean effect. The effect sizes for each study were equal to the mean differences between affected and control groups, standardized by a pooled standard deviation. Due to this standardization, data was able to be integrated across platforms. A common parameter for inter-study variability was incorporated into the model, and statistical significance was determined by permutation tests. Parmigiani *et al.* [20] introduced an integrative correlation approach to combining data from multiple platforms. This procedure evaluated gene expression consistencies across platforms rather than pooling gene expression values. Using lung cancer data, this method identified genes with reproducible expression patterns across studies and improved correlation across studies. Additional theoretical approaches for combining data from different platforms include adding covariates to models to account for the differences among data types [24,25], although this has not been applied in a microarray setting. While the studies of Rhodes *et al.* [17], Choi *et al.* [18], Parmigiani *et al.* [20] and others provide methods for integrating data across platforms, other authors have shown the difficulties in such an approach ([26,27]; discussions in [25,28]). Working with cell lines, Kuo *et al.* [26] and Jarvinen *et al.* [27] both conclude that combining data across platforms is unreliable. Due to these difficulties, other meta-analysis methods focus on incorporating data from one platform only [23,24]. Here, we focus our approach to combine microarray data from the same platform, cDNA microarrays, and assume that the data has either been pre-normalized across studies or that there are no outlying studies.

Choi *et al.* [18] also provided an alternative Bayesian meta-analysis method to their random effects approach. In this Bayesian model, uninformative prior distributions were assigned to the overall mean effects and the inter-study variation parameter. Within-study gene effects were modelled as *t*-distributions, and posterior estimates of the overall mean effect for each gene were produced by smoothing effects across studies. The authors demonstrated that Bayesian meta-analysis is more robust and flexible than traditional methods, confirming the findings of DuMouchel and Harris [29]. Bayesian models are also well-suited to data with many levels of replication, including replicate slides within repeated identical experiments. Due to these advantages, we introduce a Bayesian hierarchical model that provides a principled framework for incorporating data from multiple independent cDNA microarray studies with several sources of replication. Unlike the approach of Choi *et al.* [18], which smoothes the gene effects into one average, our method produces the posterior probability of differential expression based on gene expression levels across studies. Thus, inter-study variability does not need to be estimated by our model. The probability of differential expression provides a direct method for ranking genes, and also for estimating both integration-driven discovery rates and false discovery rates. In simulations, we illustrate that pooling studies increases the number of discovered genes for given thresholds of probabilities of differential expression and false discovery rates, compared to individual studies. In addition, for a fixed top number of genes, the pooled model identifies considerably more differentially expressed genes than separate studies. We also illustrate our method using experimental data from two independent studies in *Bacillus (B.) subtilis*.

### *Background on cDNA microarray experiments*

cDNA microarrays measure the amount of messenger RNA (mRNA) contained in an experimental sample. They are produced by robotic arrayers, which place entire gene sequences complementary to mRNA onto glass slides. In an experiment, the mRNA in two samples, e.g. treated and control, are fluorescently labeled with two different dyes, typically red and green (Cy5 and Cy3), and mixed together. The combined sample is hybridized to the array, and complementary sequences bind to each other. The

relative amounts of mRNA present in the two samples are measured by scanning the slide with two different wavelengths. The resulting fluorescent intensity values for the red and green-labeled mRNA are then compared by using the ratio of intensities. For further details, see [30-33]. We use log-ratios of intensities in each study since they even out highly skewed distributions and give a more realistic sense of variation [34].

## Results and discussion

### *Bayesian model for pooling multiple studies*

Biologists often conduct multiple but different studies that all target the same biological system or pathway. For example, when studying the effect of a key transcription factor $\sigma^E$ in *B. subtilis*, Eichenberger *et al.* [35] conducted both the $\sigma^E$ knockout and the $\sigma^E$ over-expression experiments (i.e. the mutant and induction experiments, respectively; see Methods). Thus, those genes that are up-regulated in one experiment should be down-regulated in another. Pooling both experiments can help more accurately identify true target genes. More generally, we may imagine having available multiple independent studies of one specific biological system. We assume that each study contains only two conditions: an experimental and a control. It is desirable to combine information from these studies in a principled way. Our model to achieve this goal is as follows:

$$\begin{aligned}
y_{jgse} \mid \mu_{jge} &\sim \mathrm{N}\left(\mu_{jge}, \tau_{jg}^2\right), j = 1,\ldots,J; g = 1,\ldots,G; e = 1,\ldots,E; s = 1,\ldots,S_e \\
\mu_{jge} \mid \theta_{jg} &\sim \mathrm{N}\left(\theta_{jg}, \sigma_{jg}^2\right), j = 1,\ldots,J; g = 1,\ldots,G; e = 1,\ldots,E \\
\theta_{jg} \mid I_g = 0 &\sim \mathrm{N}\left(0, \eta_{jg0}^2\right) \\
\theta_{jg} \mid I_g = 1 &\sim \mathrm{N}\left(0, c_j \times \eta_{jg0}^2\right) \\
I_g &\sim \text{Bernoulli}(p) \\
p &\sim \text{Uniform}(0,1),
\end{aligned} \tag{1}$$

for $j = 1,\ldots, J$ independent studies. Here, $y_{jgse}$ is the microarray data, i.e. the normalized log-expression ratios for gene $g$, experiment $e$, slide $s$, and $\mu_{jge}$ is the average over all slides $S_e$ within experiment $e$ of study $j$. $\theta_{jg}$ is the log-expression ratio for each gene of study $j$. Conjugate inverse chi-squared prior distributions are assigned to $\tau_{jg}^2$ and $\sigma_{jg}^2$, for which we use the notation $\tau_{jg}^2 \sim k\tilde{\tau}_j^2 / \chi_k^2$ and $\sigma_{jg}^2 \sim h\tilde{\sigma}_j^2 / \chi_h^2$. Here, $(\chi_k^2)^{-1}$ denotes the standard inverse $\chi^2$ distribution with $k$ degrees of freedom, and $\tilde{\tau}_j^2$ and $\tilde{\sigma}_j^2$ are scale parameters of the inverse chi-squared distribution and are derived from the data. The parameter $\tilde{\tau}_j^2$ is equal to slide variation, $\tilde{\sigma}_j^2$ is equal to experiment variation for study $j$, and the degrees of freedom $h$, $k$ are assumed known. We define $I_g \sim$Bernoulli($p$) as the indicator variable for differential expression of gene $g$, i.e. $\theta_{jg} \neq 0$, $j = 1,\ldots, J$, where $p$ is the percent of differentially expressed genes. Thus, Prob($I_g = 1$) = $p$, where

$$I_g = \begin{cases} 0 \text{ if } \theta_{jg} = 0, & j = 1,\ldots,J \\ 1 \text{ if } \theta_{jg} \neq 0, & j = 1,\ldots,J \end{cases}$$

Here, genes are divided into two groups, non-expressed ($I_g = 0$) and expressed ($I_g = 1$), with respective probabilities (1-$p$) and $p$. The model produces the posterior distribution for $D_g$ = Prob($I_g = 1$|data), which is the basis for inference. For prior distributions, when $I_g = 0$, we assume the $\theta_{jg}$ are distributed normally with mean zero and small variance $\eta_{jg0}^2$; when $I_g = 1$, we assume the $\theta_{jg}$ are distributed normally with mean zero and large variance $c \times \eta_{jg0}^2$. A Markov chain Monte Carlo (MCMC) implementation of the model [36] simulates posterior distributions for each parameter. See Methods for more details on the prior distributions and the MCMC implementation. For each gene, we calculate the posterior probability $D_g$ of differential expression over all studies, and rank the genes based on $D_g$. The prior estimates of the variance parameters $\tau_{jg}^2$ and $\sigma_{jg}^2$ are similar to Tseng *et al.* [2]. Our prior structure for a single experiment is similar to Gottardo *et al.* [8] except that we place a Uniform prior distribution on $p$ rather than estimating $p$ through an iterative algorithm. We also have one more level of variation than the model of Gottardo *et al.* [8], i.e. variation over slides within experiments. Our underlying hierarchical Gaussian model is also similar to the BAM models of Ishwaran and Rao [14,15], except that the BAM models are designed for a two-sample problem, while our model assumes that the data are ratios of treatment and control intensities. We evaluate our model using false discovery rates and integration-driven discovery rates, defined in the following.

### *Integration-driven discovery*

Choi *et al.* [18] define the integration-driven discovery rate (IDR) as the number of genes discovered in a meta-analysis that were not discovered in any of the individual studies alone, divided by the total number of discoveries. IDR represents the gain in information from combining studies versus individual studies. For our model, we fix a threshold value, $\gamma$, and label genes differentially expressed if ($D_g \geq \gamma$). The IDR is defined as the number of genes that

**Table 1: Results for two-study simulations. Integration-driven discovery rate (IDR) and the number of discovered genes for various threshold values of the posterior probability of differential expression, $\gamma$, and three simulated levels of the percent of differentially expressed genes *p* = 5%, 10%, 25%. The true false discovery rate (tFDR) is controlled at 5% for all pooled studies.**

| Simulated *p* | | | | | |
|---|---|---|---|---|---|
| **0.05** | $\gamma$ | 0.50 | 0.90 | 0.95 | 0.99 |
| | **IDR** | 4.6% | 22.1% | 33.3% | 42.9% |
| | **# Discovered genes** | 109 | 86 | 78 | 70 |
| **0.10** | $\gamma$ | 0.50 | 0.90 | 0.95 | 0.99 |
| | **IDR** | 3.5% | 12.6% | 14.3% | 29.8% |
| | **# Discovered genes** | 231 | 191 | 175 | 161 |
| **0.25** | $\gamma$ | 0.50 | 0.90 | 0.95 | 0.99 |
| | **IDR** | 1.9% | 3.2% | 8.4% | 20.3% |
| | **# Discovered genes** | 642 | 528 | 499 | 434 |

are labelled differentially expressed in the pooled analysis and are not differentially expressed in any of the individual studies:

$$\mathrm{IDR}(\gamma) = \frac{\#\ \text{genes}[(D_g \geq \gamma)\ \text{in pooled analysis}]\ \text{and}\ [(D_g < \gamma)\ \text{in all individual studies}]}{\#\ \text{genes}[(D_g \geq \gamma)\ \text{in pooled analysis}]}.$$

### *False discovery rate*

Benjamini and Hochberg [37] introduced the false discovery rate (FDR), which is defined as the number of false discoveries divided by the number of discoveries. We refer to this as the true false discovery rate (*t*FDR), which can be exactly computed in our simulation studies since we know which genes are truly differentially expressed. Further applications of FDR to microarrays include [38-40]. Genovese and Wasserman [41] define the posterior expected FDR (*pe*FDR) as:

$$pe\mathrm{FDR} = E\left(\mathrm{FDR} \mid \boldsymbol{Y}\right) = \frac{\sum_g \left(1 - D_g\right)\delta_g}{\sum_g \delta_g},$$

with $\delta_g$ an indicator for differentially expressed genes (see also Do *et al.* [13]). In the simulated data, we use *t*FDR and compare *t*FDR to *pe*FDR; for the experimental data, we have no choice but to use *pe*FDR.

### *Simulation results for two studies*

We simulated data for two studies, Study 1 and Study 2, with similar format to the *B. subtilis* mutant and induction studies, using three different values for the percent of truly differentially expressed genes: *p* = 5%, 10% and 25% (see Methods). We implemented Model (1) for each of the two simulation studies separately and in a pooled analysis. The IDR ranged from 1.9% to 42.9% for all values of *p* for $\gamma \geq 50\%$, with maximum *t*FDR of 5% for the pooled analysis (Table 1). Note that *t*FDR is low for large values of $\gamma$ due to the simulation procedure. The IDR increases as $\gamma$ increases. IDR is also smaller for larger values of *p* (Figure 1a); this is due to the larger variability between studies. As a result, fewer genes have $D_g$ less than $\gamma$ in both studies separately, which reduces IDR. In addition to identifying highly expressed genes by choosing a $\gamma$ threshold, researchers often choose a maximum *t*FDR and examine lists of differentially expressed genes with corresponding *t*FDR. In Figure 1b, we display all *t*FDR levels < 20% for *p* = 10% and show the number of discovered genes for the two individual studies and the pooled analysis. This plot shows the considerable increase in the number of differentially expressed genes found in the pooled analysis versus the separate analyses for the same level of *t*FDR.

In addition to choosing a threshold value of $D_g$ or FDR, researchers are often interested in the top set of genes only, i.e. the top 300 genes. For this reason, we rank the genes based on $D_g$ in both the pooled and individual analyses and compare the resulting numbers of differentially expressed genes that are included in the top genes. For each of the three simulation studies, *p* = 5%, 10%, 25%, we choose a threshold of the top *p*% of genes. We find that the pooled model always identifies a larger number of differentially expressed genes than individual studies (Table 2).

We also compared *pe*FDR to *t*FDR for the simulation data to ensure that our *pe*FDR is a reasonable approximation to the true values. As seen from Figure 2, which displays all values of *pe*FDR versus *t*FDR for the simulation results, the *pe*FDR was always larger than *t*FDR, so that *pe*FDR is a conservative estimate of *t*FDR. The average differences between *pe*FDR and *t*FDR were less than 2.3% for all pooled simulation results. The maximum difference between *pe*FDR and *t*FDR decreased as the simulated percent of truly differentially expressed genes *p* increased. Specifically, for *p* = 5%, the average difference between *pe*FDR and *t*FDR was 2.3%, with maximum difference of

**Table 2: Number of differentially expressed (D.E.) genes for fixed top numbers of genes. The number of differentially expressed genes discovered by the pooled model and individual models for fixed threshold numbers of top genes, including the two-study simulation model, *p* = 5%, 10%, 25%, and the five-study simulation model, *p* = 10%.**

| Two-study simulation data | Threshold number of genes | # of D.E. genes, pooled model | # of D.E. genes, individual study 1 | # of D.E. genes, individual study 2 |
|---|---|---|---|---|
| *p* = 5% | 150 | 128 | 97 | 108 |
| *p* = 10% | 300 | 261 | 211 | 218 |
| *p* = 25% | 750 | 669 | 569 | 572 |

| Five-study | Threshold | # of D.E. genes, | # of D.E. genes, individual study number | | | | |
|---|---|---|---|---|---|---|---|
| | | | 1 | 2 | 3 | 4 | 5 |
| *p* = 10% | 300 | 278 | 211 | 218 | 229 | 221 | 216 |

12.6% at *t*FDR = 19.9%. For *p* = 10%, the average difference was 2.2%, with maximum difference of 7.8% at *t*FDR = 3.1%. For *p* = 25%, the average difference was 2.3%, with maximum difference of 5.1% at *t*FDR = 23.8%.

### *Simulation results for five studies*

We also assessed our model for a meta-analysis with five studies. For this, we used the same Study 1 and Study 2 as in the previous section, and *p* = 10%. We then simulated three further studies similar to Study 1, but with different model parameters for $\eta_{jg0}^2, c$, and variation over slides and experiments (see Methods). The IDR was 7.1% for $\gamma$ = 0.95, and 12.8% for $\gamma$ = 99%, with *t*FDR of 0% in the pooled analysis for these levels of $\gamma$. The IDR was lower for the same levels of $\gamma$ for the five-study versus two-study pooled analysis. This was due to the larger variation between the five studies, resulting in fewer genes with $D_g$ less than $\gamma$ in all studies, which reduced IDR. We plot IDR versus $\gamma$ in Figure 3a. Figure 3b displays the number of discovered genes for the pooled analysis versus the five separate analyses for *t*FDR < 20%, which again shows a considerable increase for the pooled analysis.

We also show the number of differentially expressed genes identified by the pooled versus individual analyses for a fixed value of top expressed genes in Table 2. For the top 300 genes, the pooled model again identifies more differentially expressed genes than individual studies. We also compared *pe*FDR to *t*FDR in Figure 2d. The average difference was 0.54%, with maximum difference of 2.7%at *t*FDR = 0.36%. These values are smaller than the results for the two-study simulations, showing improved accuracy of *pe*FDR when pooling more data.

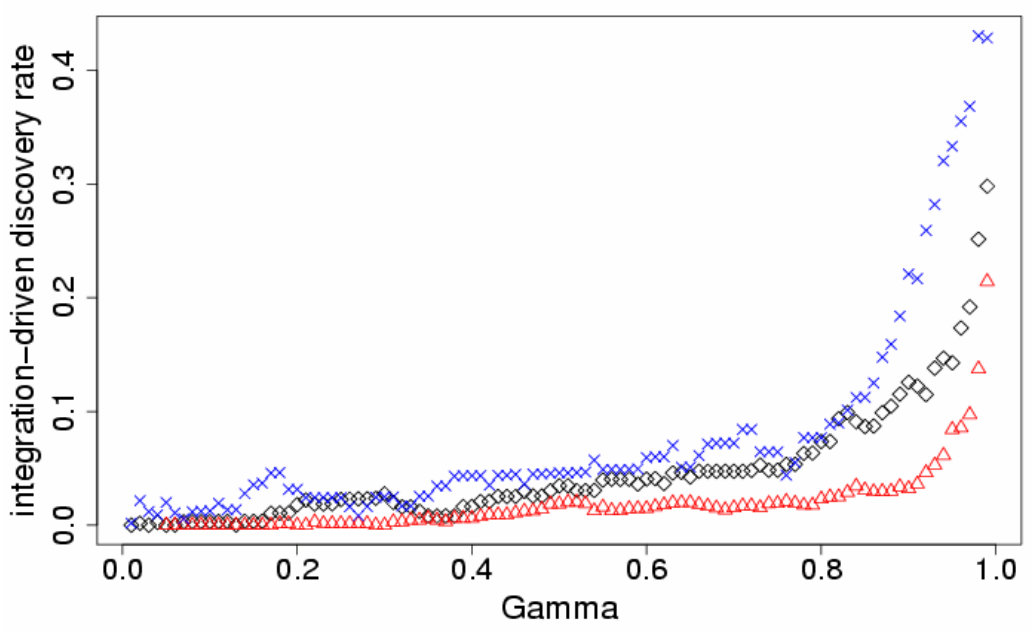
### *Experimental data results*

We implemented Model (1) to pool the mutant and induction *B. subtilis* studies, with 2,515 genes that had expression in both studies. We also implemented Model (1) for each study individually. For values of $\gamma \geq$ 50% and maximum *pe*FDR of 11.5% for the pooled analysis, the IDR ranged from 8.2% to 53.3% (Table 3). We plot IDR versus $\gamma$ in Figure 4a. Figure 4b presents the number of discovered genes for *pe*FDR < 20% for both the separate and pooled analyses. The induction study had much lower log-ratios of expression than the mutant study; the average 97.5%-ile for the induction experiments was 0.65 versus 1.59 for the mutant experiments. As a result, the maximum $D_g$ value for the induction study was 0.84, with minimum corresponding *pe*FDR of 16%. In contrast, the mutant study had 33 genes with $D_g$ of 1.0. Even though the values of $D_g$ were lower for the induction than the mutant study, we found that combining the two data sets resulted in more discoveries of differentially expressed genes than either study alone for fixed levels of *pe*FDR.

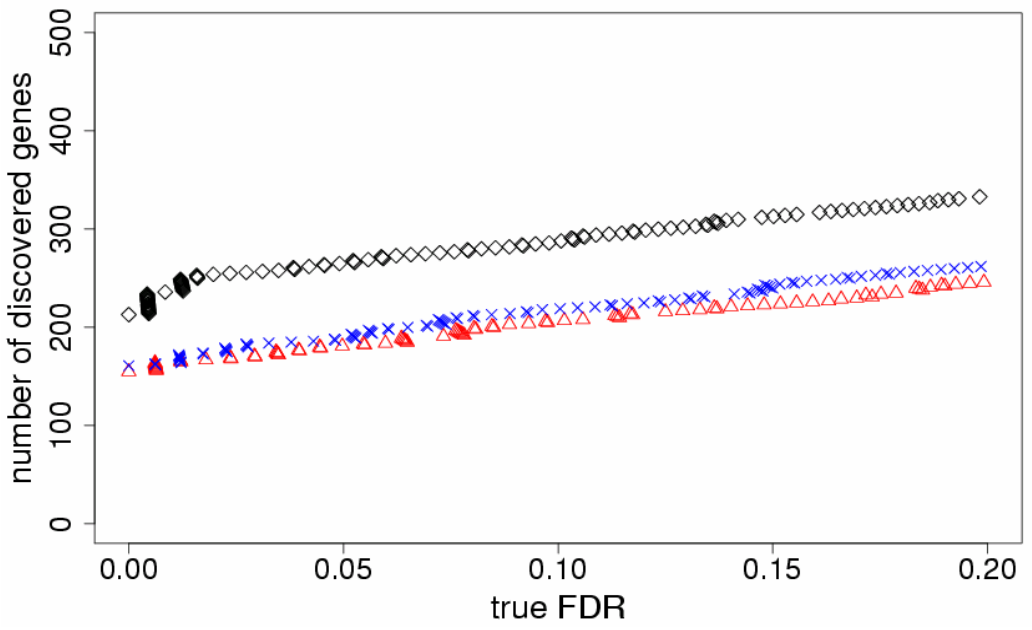
## Conclusion

We demonstrated here the usefulness of a Bayesian hierarchical model for pooling data across independent microarray studies with several sources of variation. The pooled method provides a systematic analysis framework, producing probability estimates of differential expression for each gene. These estimates are used to rank genes, calculate IDR, and produce posterior expected FDR values.

In the simulation of two and five studies, we found an appreciable increase in the IDR for various thresholds of the probability of differential expression, with corresponding low levels of *t*FDR. When fixing *t*FDR, we found more genes discovered in the pooled analysis than the separate analyses. When setting a threshold for the top genes of interest, the pooled model identified more truly




(a)




(b)

**Figure 1**

**IDR and discovered genes versus *t*FDR for the two-study simulation data**. a) Integration-driven discovery rate (IDR) versus threshold values of posterior probabilities of differential expression, $\gamma$, for the two-study simulated data and percent of differentially expressed genes $p$ = 5% (blue checks), 10% (black diamonds), 25% (red triangles); b) The maximum number of differentially expressed genes versus true false discovery rate (*t*FDR) for individual analyses of Study 1 (red triangles), Study 2 (blue checks) and pooled analysis (black diamonds), for two-study simulated data and percent of differentially expressed genes $p$ = 10%.

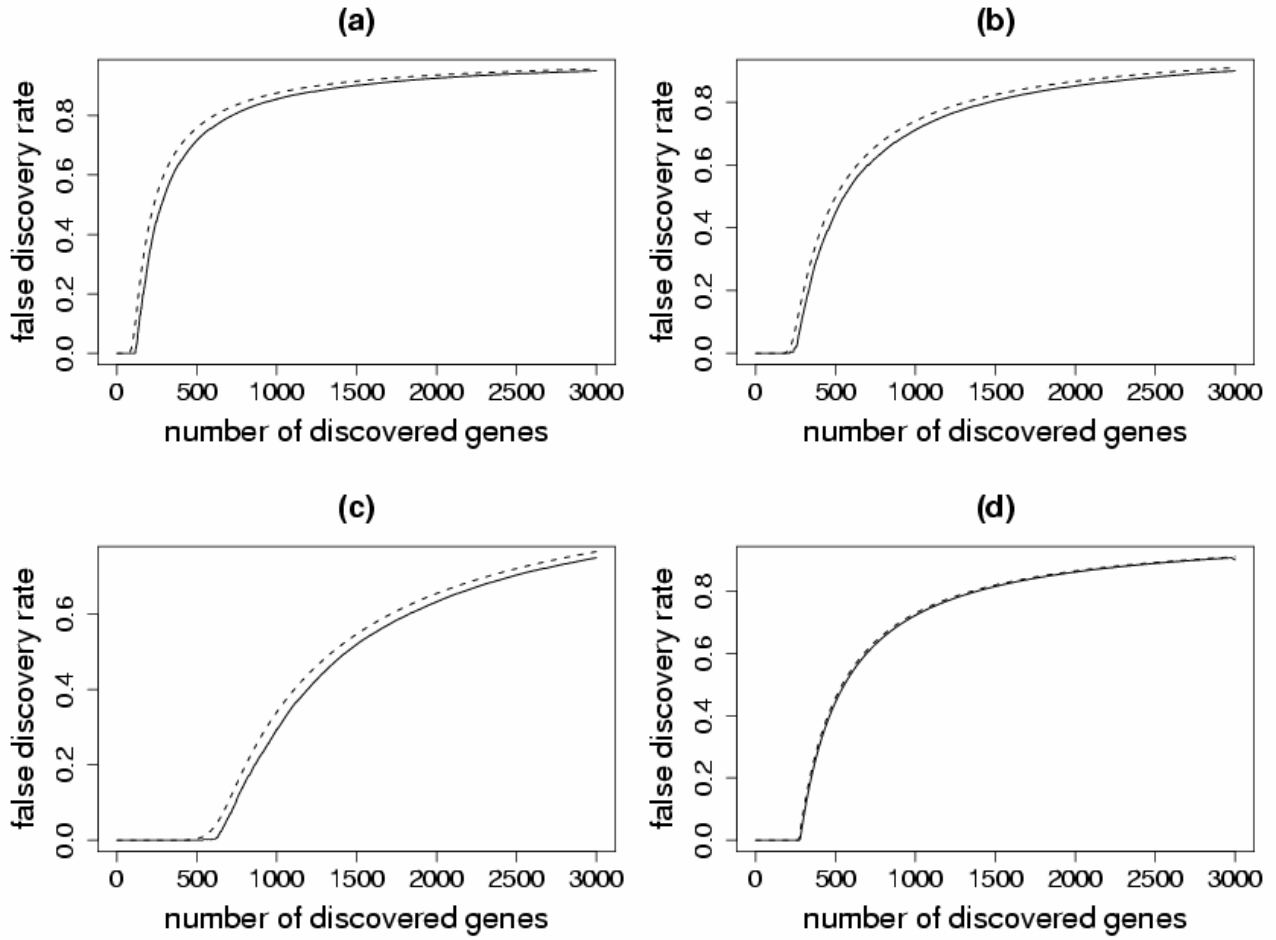



**Figure 2**
**True false discovery rate versus posterior expected false discovery rate for the simulation data**. True false discovery rate (*t*FDR) (solid lines) and posterior expected false discovery rate (*pe*FDR) (dashed lines) versus the number of discovered genes for: a) two-study simulation data, $p$ = 5%; b) two-study simulation data, $p$ = 10%; c) two-study simulation data, $p$ = 25%; d) five-study simulation data, $p$ = 10%.

differentially expressed genes than individual analyses. In the simulation of five studies, the IDR was somewhat lower than for two studies, but was still considerable. When comparing the *pe*FDR to *t*FDR in simulations, we found reasonable agreement, with *pe*FDR overestimating *t*FDR on average by less than 3%. The difference between *pe*FDR and *t*FDR decreased for the simulation of five studies, indicating that pooling more data improves the posterior estimation of FDR. In our analysis of experimental data, the IDR was also large. One study had somewhat lower probabilities of differential expression, which resulted in more discoveries when the data was pooled. We conclude that combining information across studies strengthens the probabilities of differential expression, improves IDR, and increases the number of discovered genes for fixed *t*FDR, *pe*FDR and fixed top percent of genes than individual study analyses.

Our model is designed for studies from the same platform. In the *B. subtilis* experimental data, a common control sample was used for the mutant and induction studies. However, our model does not require a common reference sample across studies, and assumes the studies are independent. In addition, all studies do not need to have the same array-design layout. This is due to the studies being linked only through the common parameter of differential expression, $p$; no other parameters are shared between studies. For example, one study could have only

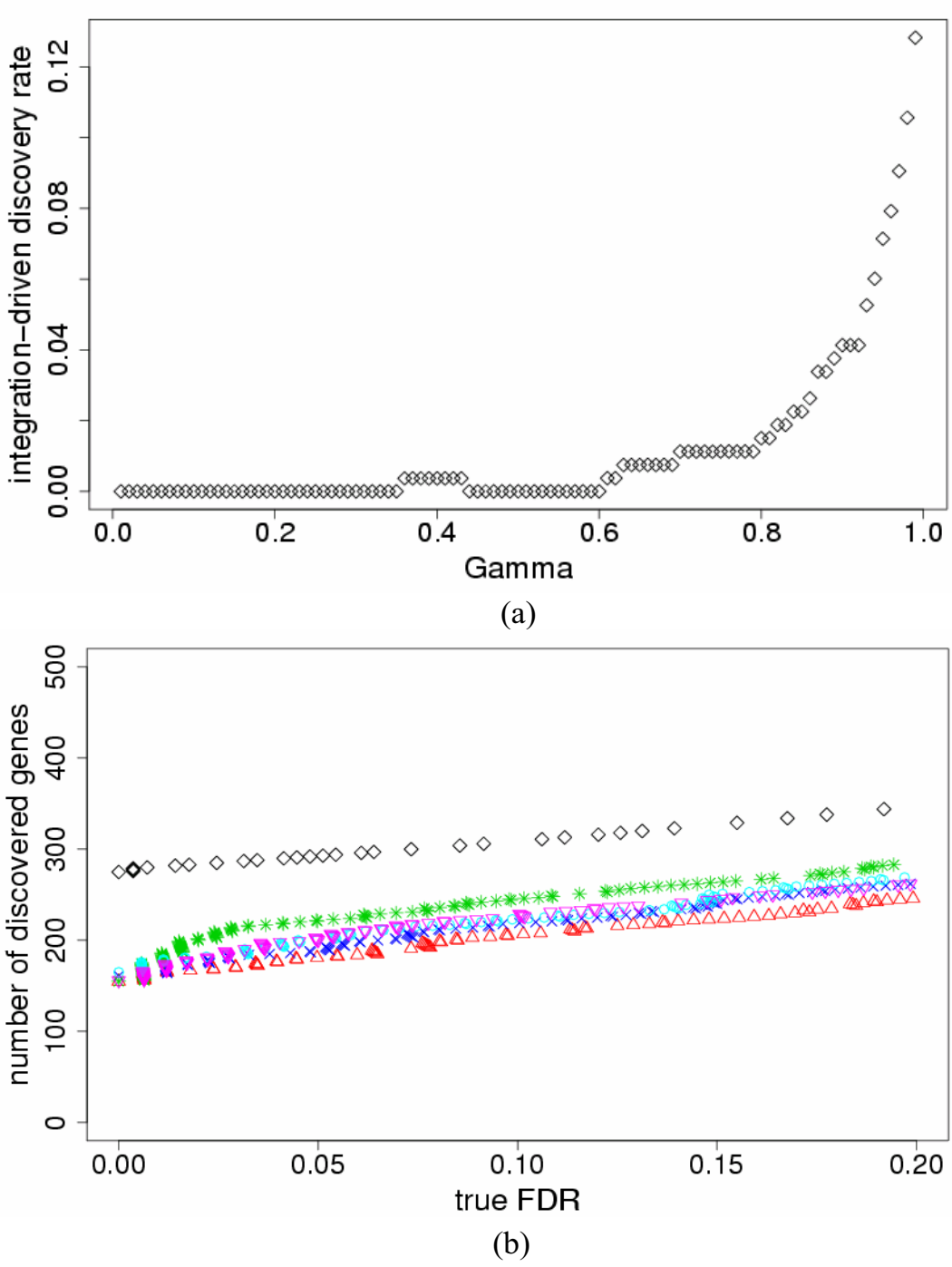



**Figure 3**
**IDR and discovered genes versus *t*FDR for the five-study simulation data**. a) Integration-driven discovery rate (IDR) versus threshold values of posterior probabilities of differential expression, $\gamma$, for the five-study simulated data and percent of differentially expressed genes $p$ = 10%; b) The maximum number of differentially expressed genes versus true false discovery rate (*t*FDR) for individual analyses of Study 1 (red triangles), Study 2 (blue checks), Study 3 (green stars), Study 4 (turquoise circles), Study 5 (pink inverted triangles) and pooled analysis (black diamonds), for five-study simulated data and percent of differentially expressed genes $p$ = 10%.

**Table 3: Results for *Bacillus subtilis* experimental data. Integration-driven discovery rate (IDR), posterior expected false discovery rate (*pe*FDR) and the number of discovered genes for various threshold values of the posterior probability of differential expression, $\gamma$, for the pooled analysis of the *B. subtilis* mutant and induction experimental study data.**

| $\gamma$ | IDR | Posterior expected FDR | # Discovered genes |
|---|---|---|---|
| 0.995 | 53.3% | 0.001 | 89 |
| 0.99 | 50.0% | 0.002 | 96 |
| 0.95 | 28.5% | 0.013 | 130 |
| 0.90 | 20.1% | 0.025 | 144 |
| 0.50 | 8.2% | 0.115 | 194 |

replicate slides, and another study could have both replicate slides and replicate experiments. We also assume that there are either no outlying studies or that the data has been scaled across studies before analysis. Future work will address the issues of pooling studies from different platforms and sets of studies that may contain outliers.

## Methods

### *Simulation data for two studies*

We simulated data for two studies, with the same format as the *B. subtilis* mutant and induction experimental data (see Methods: Experimental data), with the percent of differentially expressed genes of $p$ = 5%, 10% and 25%. Each study had 3,000 genes; the first study had 5 replicate slides within 3 replicate experiments, and the second study had 4 replicate slides within 3 replicate experiments. We simulated data from Model (1), with parameters similar to those found in the experimental data. For Study 1, we used $\eta^2_{jg0}$ = 0.015, and $c$ = 66.67. The variance across slides was set to 0.074, and across experiments to 0.029. For Study 2, we used $\eta^2_{jg0}$ = 0.02, and $c$ = 40. The variance across slides was set to 0.02, and across experiments to 0.026.

Between study variance for the experimental data was 0.067 for all 2,515 genes, and 0.296 for the top 10% of genes. The simulated data had similar between study variance for all genes, and higher variability for the top genes than the experimental data. The between study variance was 0.053 for $p$ = 5%, 0.105 for $p$ = 10% and 0.23 for $p$ = 25% for all 3,000 genes, with between study variance for the top genes of 0.714 for $p$ = 5%, 0.887 for $p$ = 10% and 0.86 for $p$ = 25%. For each gene, log-expression ratios are simulated from normal distributions, independently of other genes. Although expression is expected to have some correlation among genes, this is difficult to model, and we thus assume independence for simulation purposes. The independence assumption was also used in simulation studies by other authors (see, for example, [8,9]).

### *Simulation data for five studies*

For the simulation of five studies, we used the Study 1 and Study 2 data from the previous section, and simulated data for 3 additional studies, with $p$ = 10% for all studies. For Studies 3, 4 and 5, we simulated 5 replicate slides within 3 replicate experiments. For the parameters $\eta^2_{jg0}$ ,$c$ and variance across slides and experiments, we used a range of values that were either between the values for Study 1 and Study 2, or somewhat larger or smaller than these two studies. For the Study 3, we used $\eta^2_{jg0}$ = 0.017, and $c$ = 70.6. The variance across slides was set to 0.05, and across experiments to 0.02. For Study 4, we used $\eta^2_{jg0}$ = 0.02, and $c$ = 55. The variance across slides was set to 0.04, and across experiments to 0.022. For Study 5, we used $\eta^2_{jg0}$ = 0.015, and $c$ = 60. The variance across slides was set to 0.06, and across experiments to 0.03. Between study variance ranged from 0.098 to 0.153 for all 5 studies for all 3,000 genes, and from 0.827 to 1.35 for the top 10% of genes, which was higher than the two-study experimental data.

### *Experimental data*

The *B. subtilis* experiments were designed to identify sporulation genes under the control of sigma factor E ($\sigma^E$). Two complementary experimental setups were used, the first was a deletion of $\sigma^E$ (mutant study) and the second an overexpression of $\sigma^E$ (induction study), described in the following (for additional details, see [35,42]).

#### *Mutant study*

In the mutant study, the treated sample contained sporulating cells with a null mutation in the gene for $\sigma^E$ (i.e. the mutant sample), and the control sample contained sporulating cells that were wild type for $\sigma^E$. The wild-type/mutant ratios were examined; up-regulated genes were identified as belonging to the $\sigma^E$ regulon. In total, five microarrays were produced from three independent identical experiments; the first experiment had three replicate arrays and the second and third experiments each had one array. The number of genes spotted on the five arrays ranged from 4,268 to 4,751; these values are larger than the *B. subtilis* genome size of 4,106 due to multiple spotting of selected genes on various arrays. The percent of low quality spots that were removed from analysis ranged from 18.6% to 64.5% of values across the five arrays. In total, there were 3,713 genes with measurable expression

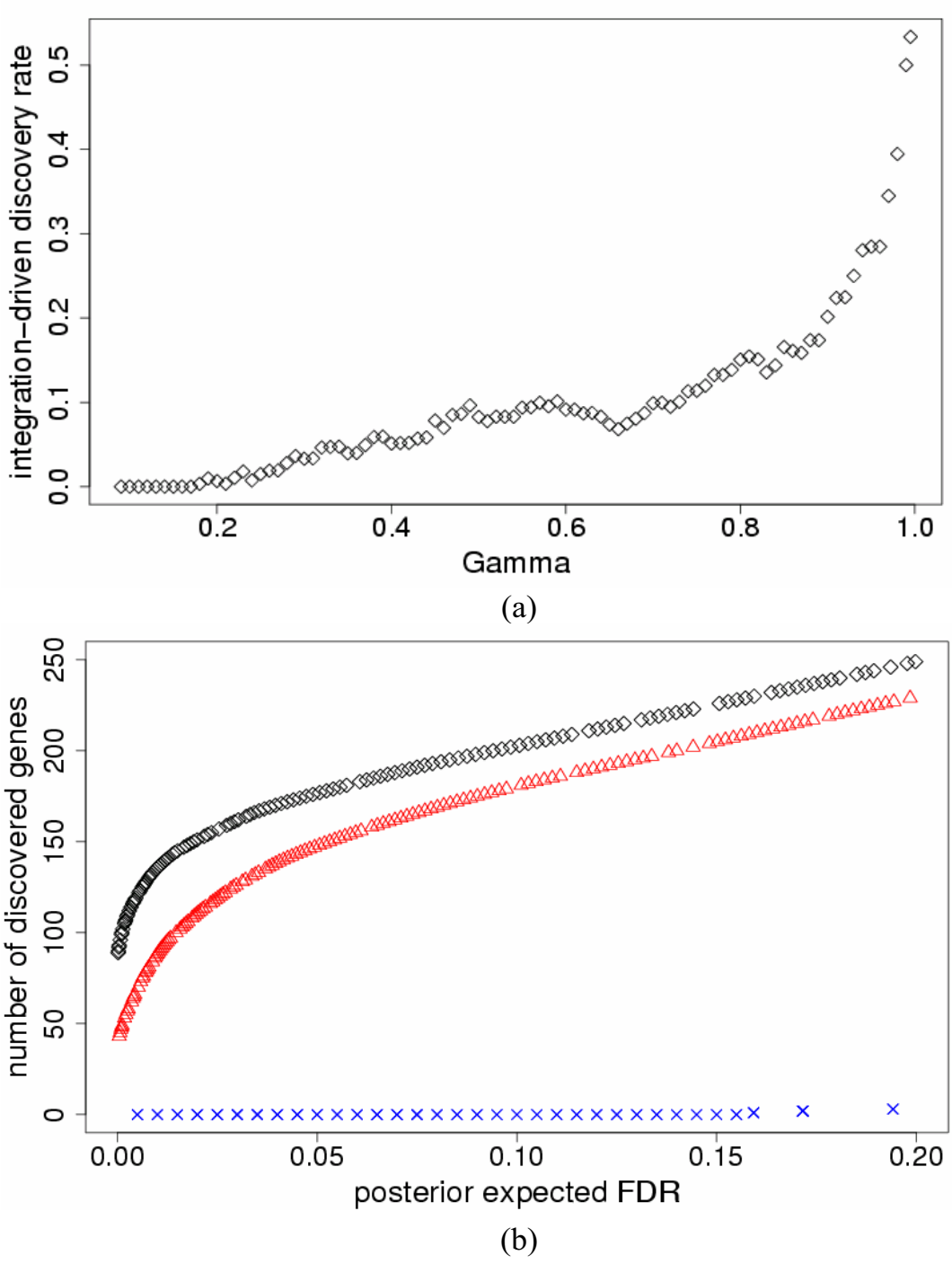



**Figure 4**
**IDR and discovered genes versus *pe*FDR for the experimental data**. a) Integration-driven discovery rate (IDR) versus threshold values of posterior probabilities of differential expression, $\gamma$, for the *B. subtilis* mutant and induction experimental study data; b) The maximum number of differentially expressed genes versus posterior expected false discovery rate (*pe*FDR) for individual analyses of the *B. subtilis* mutant study (red triangles), induction study (blue checks) and pooled analysis (black diamonds).

ratios in at least one microarray. Here, we analyze values after normalization using a rank-invariant method [2,43].

*Induction study*

In the induction study, $\sigma^E$ was overexpressed in response to an inducer, i.e. cells that had been treated with an inducer were compared to control cells. The induction/wild-type ratios were examined; up-regulated genes were identified as belonging to the $\sigma^E$ regulon. In total, four microarrays were produced from three independent identical experiments. The first two experiments each had one array, and the third experiment had two replicate arrays. The number of genes spotted on the four arrays ranged from 4,608 to 4,751; the percentage of genes detected on the arrays ranged from 33.0% to 47.4%. In total, there were 2,552 genes with measurable expression ratios in at least one microarray. Here, we again analyze the post-normalized values.

### *Markov chain Monte Carlo implementation*

In the Markov chain Monte Carlo analysis, the full conditionals are simulated as follows.

*Joint posterior distribution*

For the hierarchical model of (1), the joint distribution of the data and parameters is:

$$p(y_{jgse}, \mu_{jge}, \tau^2_{jg}, \sigma^2_{jg}, \theta_{jg}, I_g, p, \eta^2_{jg0}, c_j) = \prod_{j=1}^{J}\prod_{g=1}^{G}\prod_{e=1}^{E}\left\{\prod_{s=1}^{S_e} p(y_{jgse} \mid \mu_{jge}, \tau^2_{jg}) p(\mu_{jge} \mid \theta_{jg}, \sigma^2_{jg})\right\} p(\theta_{jg}, I_g \mid p, \Omega_j) p(\tau^2_{jg}) p(\sigma^2_{jg}) p(p) p(\Omega_j),$$

where $\Omega_j = (\eta^2_{jg0}, c_j)$, $j$ = study, $g$ = gene, $e$ = experiment, $s$ = slide.

*Prior distributions*

The prior distributions are specified as follow.

$$\tau^2_{jg} \sim k\tilde{\tau}^2_j / \chi^2_k$$
$$\sigma^2_{jg} \sim h\tilde{\sigma}^2_j / \chi^2_h$$

Here, $(\chi^2_k)^{-1}$ denotes the standard inverse $\chi^2$ distribution with $k$ degrees of freedom, and $\tilde{\tau}^2_j$ and $\tilde{\sigma}^2_j$ are scale parameters of the inverse chi-squared distribution derived from the data. $\tilde{\tau}^2_j$ is produced as follows:

$$\tilde{\tau}^2_j = \frac{1}{G\left(\sum S_e - 1\right)} \sum_{g=1}^{G}\sum_{e=1}^{E}\sum_{s=1}^{S_e}\left(y_{jgse} - y_{jg\cdot e}\right)^2,$$

where $y_{jg.e}$ is the average log-ratio of expression over the slides within an experiment:

$$y_{jg\cdot e} = \frac{1}{S_e}\sum_{s=1}^{S_e} y_{jgse}.$$

Similarly, the scale parameter for $\sigma^2_{jg}$ is calculated as follows:

$$\tilde{\sigma}^2_j = \frac{1}{G(E-1)} \sum_{g=1}^{G}\sum_{e=1}^{E}\left(y_{jg.e} - y_{jg..}\right)^2$$

where $y_{jg..}$ is the average log-ratio of expression over both slides and experiments. We use 3 degrees of freedom in each study for both $\tau^2_{jg}$ and $\sigma^2_{jg}$, i.e. $h = k = 3$. The prior distributions for the remaining parameters are as follow.

$$\begin{aligned}
\theta_{jg} \mid I_g = 0 &\sim \mathrm{N}(0, \eta^2_{jg0}) \\
\theta_{jg} \mid I_g = 1 &\sim \mathrm{N}(0, c_j \times \eta^2_{jg0}) \\
I_g &\sim \text{Bernoulli}(p) \\
p &\sim \text{Uniform}(0,1) \\
\eta^2_{jg0} &\sim as^2_1 / \chi^2_a \\
c_j &\sim bs^2_2 / \chi^2_b
\end{aligned}$$

We choose $a$, $s^2_1$ so that the prior mean of $\eta^2_{jg0}$ is 1 with variance 0.1. We choose $b$, $s^2_2$ so that the prior mean of $c_j$ is 100 with variance 10,000.

*Full conditional posterior distributions*

Each parameter is sampled from the full conditional posterior distributions by the following.

$$\begin{aligned}
\mu_{jge} \mid \text{rest} &\sim N\left(\frac{S_e y_{jg\cdot e}\sigma^2_{jg} + \tau^2_{jg}\theta_{jg}}{S_e\sigma^2_{jg} + \tau^2_{jg}}, \frac{\tau^2_{jg}\sigma^2_{jg}}{S_e\sigma^2_{jg} + \tau^2_{jg}}\right), \\
\tau^2_{jg} \mid \text{rest} &\sim \left\{\sum_{e=1}^{E}\sum_{s=1}^{S_e}(y_{jgse} - \mu_{jge})^2 + k\tilde{\tau}^2_j\right\} \Big/ \chi^2_{S_1+\ldots+S_E+k}, \\
\sigma^2_{jg} \mid \text{rest} &\sim \left[\sum_{e=1}^{E}(\mu_{jge} - \mu_{jg\cdot})^2 + h\tilde{\sigma}^2_j\right] \Big/ \chi^2_{E+h-1}.
\end{aligned}$$

Here, $\chi^2_{S_1+\ldots+S_E+k}$ denotes the standard inverse $\chi^2$ distribution with $(S_1 + \ldots + S_E + k)$ degrees of freedom, with scale parameter:

$$\left\{\sum_{e=1}^{E}\sum_{s=1}^{S_e}(\gamma_{jgse}-\mu_{jge})^2+k\tilde{\tau}_j^2\right\}\Big/(S_1+\ldots+S_E+k).$$

For $I_g = 0$, the following full conditionals are sampled:

$$\theta_{jg}\mid I_g=0,\ \text{rest}\ \sim N\left(\frac{E\eta_{jg0}^2\mu_{jg\cdot}}{E\eta_{jg0}^2+\sigma_{jg}^2},\frac{\sigma_g^2\eta_{jg0}^2}{E\eta_{jg0}^2+\sigma_{jg}^2}\right),$$

$$\eta_{jg0}^2\mid I_g=0,\ \text{rest}\sim\left[\frac{as_1^2+\theta_{jg}^2}{a+1}\right]\Big/\chi_{a+1}^2.$$

For $I_g = 1$, the following full conditionals are sampled:

$$\theta_{jg}\mid I_g=1,\ \text{rest}\ \sim N\left(\frac{Ec_j\eta_{jg0}^2\mu_{jg\cdot}}{Ec_j\eta_{jg0}^2+\sigma_{jg}^2},\frac{\sigma_{jg}^2c_j\eta_{jg0}^2}{Ec_j\eta_{jg0}^2+\sigma_{jg}^2}\right),$$

$$\eta_{jg0}^2\mid I_g=1,\ \text{rest}\sim\left[\frac{c_jas_1^2+\theta_{jg}^2}{c_j(a+1)}\right]\Big/\chi_{a+1}^2.$$

For all iterations, the following are sampled:

$$c_j\mid\text{rest}\ \sim\ \left[\frac{G'bs_2^2+\sum_{g=1}^{G'}\frac{\theta_{jg}^2}{\eta_{jg0}^2}}{G'(b+1)}\right]\Big/\chi_{G'(b+1)}^2,$$

$$p\mid\text{rest}\ \sim\ \text{Beta}\left(\sum_{g=1}^{G}I_g+1,G-\sum_{g=1}^{G}I_g+1\right),$$

$$I_g\ \sim\ \text{Bernoulli}(d_g),$$

$$d_g=p(I_g=1\mid\text{rest})=\frac{p\ \text{Prob}(\mathbf{Y}_g\mid I_g=1)}{p\ \text{Prob}(\mathbf{Y}_g\mid I_g=1)+(1-p)\ \text{Prob}(\mathbf{Y}_g\mid I_g=0)}.$$

Here $G$ = total number of genes $g$, $G'$ = set of genes with $I_g = 1$ in an iteration, and $\mathbf{Y}_g$ is the data from all studies. Since the full conditional posterior distributions are all closed form when conditioned on the values of $I_g$, the Gibbs sampler [36] is used to generate samples from these distributions. We used 5,000 iterations for all analyses, except for the five study simulation, which required 8,000 iterations, which was more than adequate. The calculations are implemented using the WinBUGS software [44].

## Availability and requirements

The WinBUGS code for executing the models is freely available.

Project name: BayesPoolMicro.

Project home page: http://www.math.umass.edu/~conlon/research/BayesPoolMicro/

Operating system: Windows 98 or later.

Other requirements: WinBUGS software version 1.4 or later [44].

License: free.

## Authors' contributions

EMC and JJS contributed to writing the computer code. All authors contributed to the development of the methodology and to writing the manuscript.

## Acknowledgements

We thank George Tseng, Jeffrey Townsend and John Staudenmayer for helpful discussion, and Patrick Eichenberger and the laboratory of Richard Losick for the *B. subtilis* microarray data and helpful advice. We also thank three anonymous referees for input that enhanced the manuscript. EMC was partially supported by a University of Massachusetts Healey Endowment Grant, and JSL was partially supported by the NIH Grant R01-HG02518-01, and the NSFChina Grant 10228102.

## References

1. Baldi P, Long AD: **Bayesian framework for the analysis of microarray expression data: reguralized t-test and statistical inferences of gene changes.** *Bioinformatics* 2001, **17:**509-519.
2. Tseng GC, Oh MK, Rohlin L, Liao JC, Wong WH: **Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects.** *Nucleic Acids Res* 2001, **29:**2549-2557.
3. Townsend JP, Hartl DL: **Bayesian analysis of gene expression levels: statistical quantification of relative mRNA level across multiple treatments or samples.** *Genome Biology* 2002, **3:**research0071.1-71.16.
4. Efron B, Tibshirani R, Storey JD, Tusher VG: **Empirical Bayes Analysis of a Microarray Experiment.** *Journal of the American Statistical Association* 2001, **96:**1151-1160.
5. Newton MA, Kendziorski CM, Richmond CS, Blattner FR, Tsui KW: **On Differential Variability of Expression Ratios: Improving Statistical Inference About Gene Expression Changes From Microarray Data.** *Journal of Computational Biology* 2001, **8:**37-52.
6. Ibrahim JG, Chen M-H, Gray RJ: **Bayesian Models for Gene Expression With DNA Microarray Data.** *Journal of the American Statistical Association* 2002, **97:**88-99.
7. Broët P, Richardson S, Radvanyi F: **Bayesian hierarchical model for identifying changes in gene expression from microarray experiments.** *Journal of Computational Biology* 2002, **9:**671-683.
8. Gottardo R, Pannucci JA, Kuske CR, Brettin T: **Statistical analysis of microarray data: a Bayesian approach.** *Biostatistics* 2003, **4:**597-620.
9. Lönnstedt I, Speed TP: **Replicated microarray data.** *Statistica Sinica* 2002, **12:**31-46.
10. Pan W: **A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments.** *Bioinformatics* 2002, **18:**546-554.
11. Kendziorski CM, Newton MA, Lan H, Gould MN: **On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles.** *Statistics in Medicine* 2003, **22:**3899-3914.
12. Newton MA, Noueiry A, Sarkar D, Ahlquist P: **Detecting differential gene expression with a semiparametric hierarchical mixture method.** *Biostatistics* 2004, **5:**155-176.
13. Do KA, Müller P, Tang F: **Bayesian mixture model for differential gene expression.** *Journal of the Royal Statistical Society C* 2005, **54:**627-644.

14. Ishwaran H, Rao JS: **Detecting Differentially Expressed Genes in Microarrays Using Bayesian Model Selection.** *Journal of the American Statistical Association* 2003, **98:**438-455.
15. Ishwaran H, Rao JS: **Spike and Slab Gene Selection for Multipgroup Microarray Data.** *Journal of the American Statistical Association* 2005, **100:**764-780.
16. Yang D, Zakharkin SO, Page GP, Brand JP, Edwards JW, Bartolucci AA, Allison DB: **Applications of Bayesian statistical methods in microarray data analysis.** *Am J Pharmacogenomics* 2004, **4:**53-62.
17. Rhodes DR, Barrette TR, Rubin MA, Ghosh D, Chinnaiyan AM: **Meta- analysis of microarrays: inter-study validation of gene expression profiles reveals pathway dysregulation in prostate cancer.** *Cancer Research* 2002, **62:**4427-4433.
18. Choi JK, Yu U, Kim S, Yoo OJ: **Combining multiple microarray studies and modeling inter-study variation.** *Bioinformatics* 2003:i84-i90.
19. Ghosh D, Barette TR, Rhodes D, Chinnaiyan AM: **Statistical issues and methods for meta-analysis of microarray data: a case study in prostate cancer.** *Functional & Integrative Genomics* 2003, **3:**180-188.
20. Parmigiani G, Garrett-Mayer ES, Anbazhagan R, Gabrielson E: **A cross-study comparison of gene expression studies for the molecular classification of lung cancer.** *Clinical Cancer Research* 2004, **10:**2922-2927.
21. Jiang H, Deng Y, Chen H, Tao L, Sha Q, Chen J, Tsai C, Zhang S: **Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes.** *BMC Bioinformatics* 2004, **5:**81.
22. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM: **Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression.** *Proc Natl Acad Sci USA* 2004, **101:**9309-9314.
23. Hu P, Greenwood CMT, Beyene J: **Integrative analysis of multiple gene expression profiles with quality-adjusted effect size models.** *BMC Bioinformatics* 2005, **6:**128.
24. Hedges LV, Olkin I: **Statistical Methods for Meta-Analysis.** San Diego, CA, Academic Press; 1985.
25. Stevens JR, Doerge RW: **Combining Affymetrix microarray results.** *BMC Bioinformatics* 2005, **6:**57.
26. Kuo WP, Jenssen TK, Butte AJ, Ohno-Machado L, Kohane IS: **Analysis of matched mRNA measurements from two different microarray technologies.** *Bioinformatics* 2002, **18:**405-412.
27. Jarvinen AK, Hautaniemi S, Edgren H, Auvinen P, Saarela J, Kallioniemi OP, Monni O: **Are data from different gene expression microarray platforms comparable?** *Genomics* 2004, **83:**1164-1168.
28. Hardiman G: **Microarray platforms – comparisons and contrasts.** *Pharmacogenomics* 2004, **5:**487-502.
29. DuMouchel WH, Harris JE: **Bayes methods for combining the results of cancer studies in humans and other species.** *Journal of the American Statistical Association* 1983, **78:**293-315.
30. Lockhart DJ, Winzeler EA: **Genomics, gene expression and DNA arrays.** *Nature* 2000, **405:**827-836.
31. Wu TD: **Analyzing gene expression data from DNA microarrays to identify candidate genes.** *Journal of Pathology* 2001, **195:**53-65.
32. Hardiman G: **Microarray technologies – an overview.** *Pharmacogenomics* 2002, **3:**293-297.
33. Southern EM: **DNA microarrays. History and overview.** *Methods Mol Biol* 2000, **170:**1-15.
34. Dudoit S, Yang YH, Callow MJ, Speed TP: **Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments.** *Statistica Sinica* 2002, **12:**111-139.
35. Eichenberger P, Jensen ST, Conlon EM, van Ooij C, Silvaggi J, Gonzalez-Pastor JE, Fujita M, Ben-Yehuda S, Stragier P, Liu JS, Losick R: **The sigmaE regulon and the identification of additional sporulation genes in *Bacillus subtilis*.** *Journal of Molecular Biology* 2003, **327:**945-972.
36. Liu JS: **Monte Carlo Strategies in Scientific Computing.** New York, Springer-Verlag; 2001.
37. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *Journal of the Royal Statistical Society B* 1995, **85:**289-300.
38. Tusher VG, Tibshirani R, Chu G: **Significance Analysis of Microarrays Applied to the Ionizing Radiation Response.** *Proceedings of the National Academy of Sciences USA* 2001, **98:**5116-5121.
39. Storey JD: **A Direct Approach to False Discovery Rates.** *Journal of the Royal Statistical Society B* 2002, **64:**479-498.
40. Storey JS, Tibshirani R: **SAM Thresholding and False Discovery Rates for Detecting Differential Gene Expression in DNA Microarrays.** In *The Analysis of Gene Expression Data: Methods and Software* Edited by: Parmigiani G, Garrett ES, Irizarry RA, Zeger SL. Springer, NY; 2003.
41. Genovese C, Wasserman L: **Operating characteristics and extensions of the false discovery rate procedure.** *Journal of the Royal Statistical Society B* 2002, **64:**499-518.
42. Conlon EM, Eichenberger P, Liu JS: **Determining and analyzing differentially expressed genes from cDNA microarray experiments with complementary designs.** *Journal of Multivariate Analysis* 2004, **90:**1-18.
43. Schadt EE, Li C, Ellis B, Wong WH: **Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data.** *J CellBiochem Suppl* 2001, **37:**120-125.
44. **The BUGS Project** [http://www.mrc-bsu.cam.ac.uk/bugs]